

# **INSTITUTE OF ACTUARIES OF INDIA**

## **Subject CS1B – Actuarial Statistics (Paper B)**

**February 2025**

### **INDICATIVE SOLUTION**

#### **Introduction:**

The indicative solution has been written by the Examiners with the aim of helping candidates. The solutions given are only indicative. It is realized that there could be other points as valid answers and examiners have given credit for any alternative approach or interpretation which they consider to be reasonable.

**Solution to Question 1:**

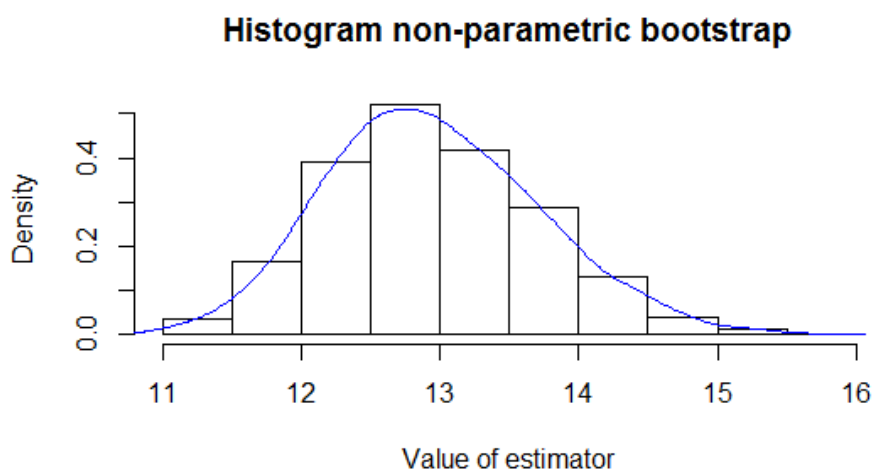
i) `data<- c(14.8, 17.6, 11.2, 13.5, 12.9, 10.8, 10.5, 12.3)` (1)

ii) `set.seed(2025)`  
`estimate<-rep(0,1000)`  
`for(i in 1:1000)`  
`{x<-sample(data,replace=TRUE);`  
`estimate[i]<-mean(x)}`

Alternate Code:

`set.seed(2025)`  
`estimate<-replicate(1000,mean(sample(data,replace=TRUE)))` (4)

iii) `hist(estimate, prob=TRUE, main="Histogram non-parametric bootstrap",xlab="Value of estimator")`  
`lines(density(estimate),col="blue")` (2)



iv) `mean(estimate)`  
`[1] 12.95271` (2)

`sd(estimate)`  
`[1] 0.7647271` (4)

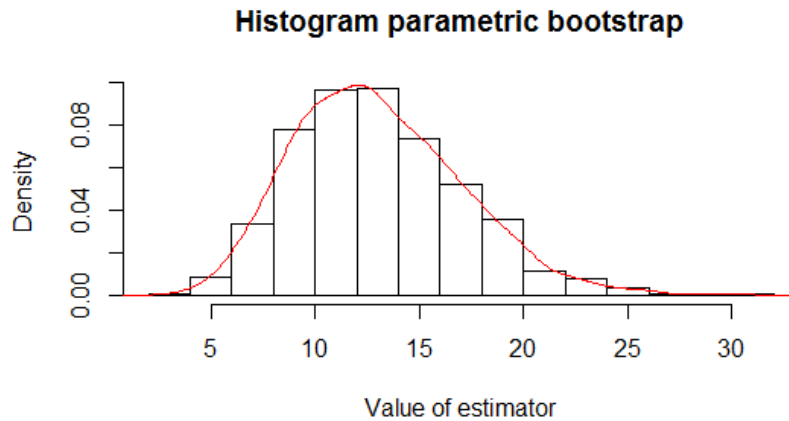
v) `set.seed(2025)`  
`param.estimate<-rep(0,1000)`  
`for(i in 1:1000)`  
`{x<-rexp(8,rate=1/mean(data));`  
`param.estimate[i]<-mean(x)}`

Alternative Code:

`set.seed(2025)`  
`param.estimate<-replicate(1000,mean(rexp(8,rate=1/mean(data))))` (4)

vi) `hist(param.estimate, prob=TRUE, main="Histogram parametric bootstrap", xlab="Value of estimator")`

`lines(density(param.estimate), col="red")` (2)



(2)

(4)

vii) `mean(param.estimate)`  
[1] 13.06449

`sd(param.estimate)`  
[1] 4.675508

(2)

viii) #a. Method of Moments Estimate

`1/mean(data)`  
[1] 0.07722008

#b. Estimate using non-parametric bootstrap

`1/mean(estimate)`  
[1] 0.07720391

#c. Estimate using parametric bootstrap

`1/mean(param.estimate)`  
[1] 0.07654337 (3)

ix) By looking at the two histograms, we can see that the empirical distribution under parametric bootstrap has a longer tail as compared to the empirical distribution under non-parametric bootstrap.

This is validated from the fact that the standard deviation of the bootstrapped samples under parametric bootstrap is higher than the standard deviation under non-parametric bootstrap.

(1)

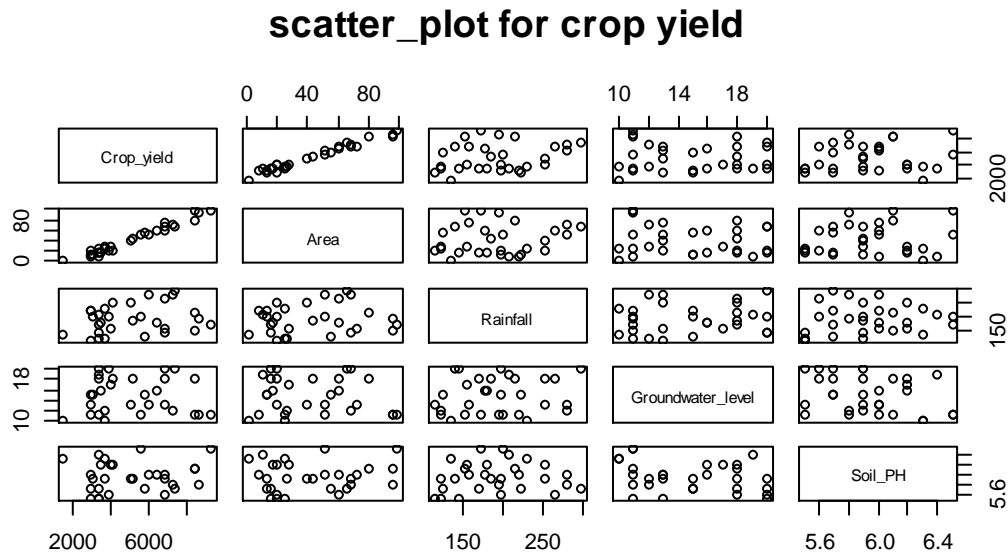
[25]

**Solution to Question 2:**

```
i) library(readr)
crop_yield <- read_csv("D:/CS1B/crop_yield.csv")

plot(crop_yield,main="scatter_plot for crop yield")
```

(1)



(1)

(2)

ii) #a

```
cor_matrix=cor(crop_yield,method="pearson")
cor_matrix
```

	Crop_yield	Area	Rainfall	Groundwater_level
Crop_yield	1.0000000	0.9999991	0.13029441	-0.1370558
Area	0.9999991	1.0000000	0.13035844	-0.1372542
Rainfall	0.1302944	0.1303584	1.0000000	0.1441443
Groundwater_level	-0.1370558	-0.1372542	0.14414434	1.0000000
Soil_PH	0.1079254	0.1075342	0.07591003	-0.2938820

```
Soil_PH
Crop_yield 0.10792539
Area       0.10753418
Rainfall   0.07591003
Groundwater_level -0.29388200
Soil_PH    1.00000000
```

(4)

b)

The relationship between crop yield and area is positive. The correlation is very strong and close to 1.

The relationship between crop yield and rainfall is positive. But the correlation coefficient is close to 0. It is quite possible that because of irrigation, the crop yield in the district is not dependent on rainfall.

The relationship between crop yield and groundwater level is negative. Correlation coefficient again is close to 0. Like rainfall it seems that because of irrigation, the crop yield is not dependent on groundwater levels.

The relationship between crop yield and soil PH is positive. Correlation coefficient is close to 0

indicating low dependence of crop yield on soil PH.

(4)

iii) #a

```
modell<-lm(crop_yield$Crop_yield~crop_yield$Rainfall)
modell
```

Call:

Call:

```
lm(formula = crop_yield$Crop_yield ~ crop_yield$Rainfall)
```

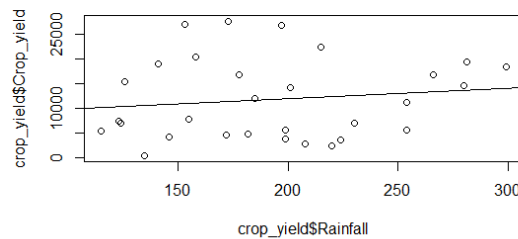
Coefficients:

```
(Intercept) crop_yield$Rainfall
7854.98      20.62
```

(3)

#b

```
plot(crop_yield$Rainfall,crop_yield$Crop_yield)abline(modell)
```



(2)

iv)  $H_0$ : Slope coefficient  $\beta = 0$  (crop yield is independent of soil PH)

$H_1$ : Slope coefficient  $\beta > 0$  (crop yield is not independent of soil PH)

(1)

```
model2<-lm(crop_yield$Crop_yield~crop_yield$Soil_PH)
confint(model2,level=0.95)
```

```
2.5 % 97.5 %
```

```
(Intercept) -70499.780 58138.45
```

```
crop_yield$Soil_PH -7773.835 13833.17
```

(2)

As the confidence interval for slope coefficient contains the value 0, we have insufficient evidence to reject the null hypothesis and hence we can say that crop yield is independent of soil PH.

(1)

(4)

v) `model3<-lm(crop_yield$Crop_yield~crop_yield$Groundwater_level)`

```
> CY_at_0 = model3$coefficients[1] + model3$coefficients[2] * 0
```

```
> CY_at_0
```

```
(Intercept)
```

```
16641.52
```

(2)

```

> avg_CY = mean(crop_yield$Crop_yield)
> avg_CY
[1] 11835.77
>
> per_chg = (CY_at_0 - avg_CY) / avg_CY * 100
> per_chg
(Intercept)
  40.60362

```

It appears that at zero groundwater level, in fact the crop yield increases by around 41% as compared to the current average crop yield. This could be mainly due to improvements in irrigation, use of technology in agriculture, etc. where despite of zero ground water levels, the crop yield seems to increase.

(2)

(4)

```

vi) > model4 = lm(crop_yield$Area ~ crop_yield$Crop_yield)
> model4

```

Call:

```
lm(formula = crop_yield$Area ~ crop_yield$Crop_yield)
```

Coefficients:

```

(Intercept) crop_yield$Crop_yield
  -0.108435      0.003572

```

```

>
> area= model4$coefficients[1] + model4$coefficients[2] * 200000
> area
(Intercept)
  714.2534

```

So, 714.2534 hectares of paddy cultivation is needed in the district in order to be self-sufficient in rice consumption.

(4)

```

vii) > #a
>
> model5=lm(crop_yield$Crop_yield~crop_yield$Area+crop_yield$Rainfall+crop_yield$Groun
dwater_level+crop_yield$Soil_PH)
> model5

```

Call:

```
lm(formula = crop_yield$Crop_yield ~ crop_yield$Area + crop_yield$Rainfall +
  crop_yield$Groundwater_level + crop_yield$Soil_PH)
```

Coefficients:

```

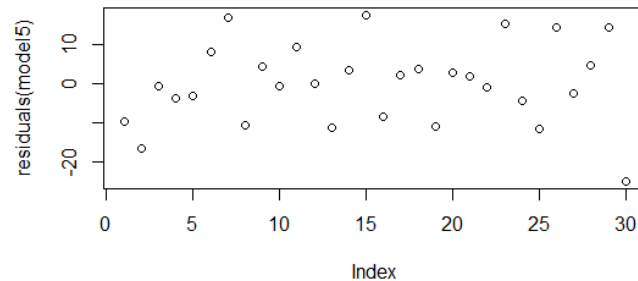
(Intercept)      crop_yield$Area
  -63.44955      279.97440
crop_yield$Rainfall crop_yield$Groundwater_level
  -0.02513      0.88383
crop_yield$Soil_PH
  14.35632

```

(2)

#b

```
plot(residuals(model5))
```



(2)

```
viii) > model6=glm(crop_yield$Crop_yield~crop_yield$Area+crop_yield$Rainfall+crop_yield$Groundwater_level+crop_yield$Soil_PH,family=Gamma(link=inverse))
> model6
```

Call: glm(formula = crop\_yield\$Crop\_yield ~ crop\_yield\$Area + crop\_yield\$Rainfall + crop\_yield\$Groundwater\_level + crop\_yield\$Soil\_PH, family = Gamma(link = inverse))

Coefficients:

(Intercept)	crop_yield\$Area
1.912e-04	-1.941e-06
crop_yield\$Rainfall	crop_yield\$Groundwater_level
-1.071e-07	-2.916e-06
crop_yield\$Soil_PH	
1.280e-05	

Degrees of Freedom: 29 Total (i.e. Null); 25 Residual

Null Deviance: 19.42

Residual Deviance: 6.705 AIC: 596.6

(2)

```
ix) > #a
>
> avg_CY
[1] 11835.77
> avg_area=mean(crop_yield$Area)
> avg_area
[1] 42.16667
> avg_rainfall=mean(crop_yield$Rainfall)
> avg_rainfall
[1] 193.1
> avg_gwl=mean(crop_yield$Groundwater_level)
> avg_gwl
[1] 14.83333
> avg_sph=mean(crop_yield$Soil_PH)
> avg_sph
[1] 5.946667
```

(2)

```

> #b
>
> exp_CY_M5 = model5$coefficients[1] + model5$coefficients[2] * avg_area + model5$coeffic
ients[3] *avg_rainfall + model5$coefficients[4] *avg_gwl +model5$coefficients[5] *avg_sph
>
> exp_CY_M5
(Intercept)
  11835.77
expected_cy_glm <- predict(model_glm, newdata = as.data.frame(t(avg_values)), type = "resp
onse")
print(paste("Expected crop yield (GLM):", expected_cy_glm))
## [1] "Expected crop yield (GLM): 8226.56266563544"

```

(4)

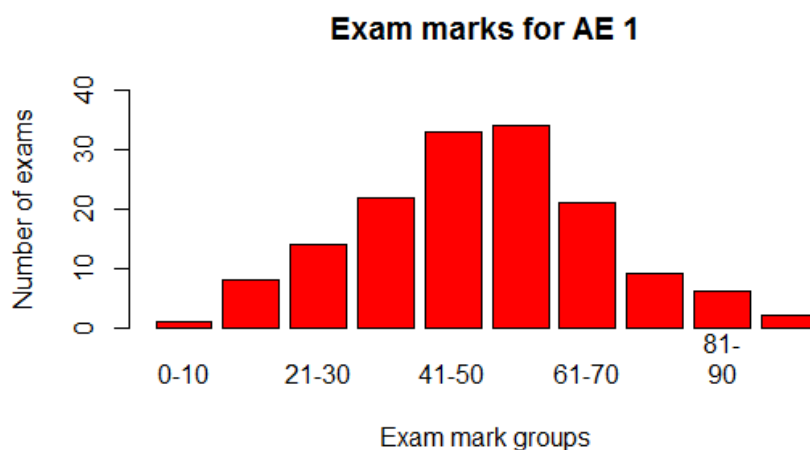
#c

It appears that the expected crop yield determined using the multiple linear regression model is very close to the average annual yield for the 30 records. However, the expected yield using the GLM Gamma Model is relatively not close to the average annual yield. Clearly the multiple linear regression model is a better fit to the data.

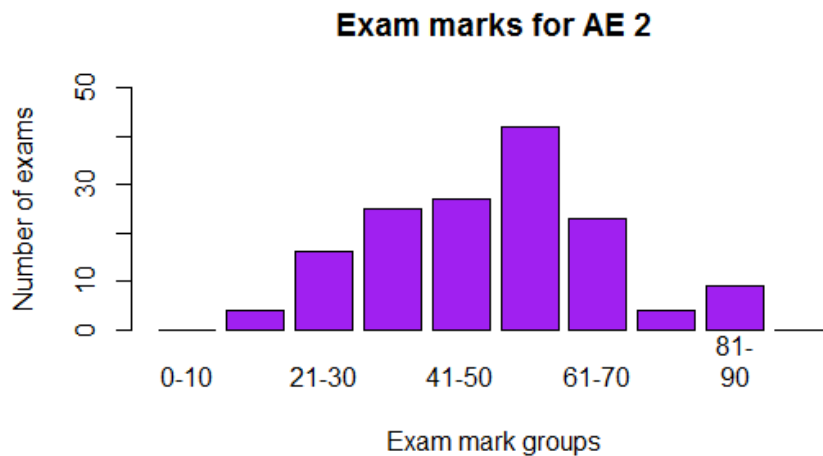
(1)  
[40]

### Solution to Question 3:

- i) `axis = c("0-10", "11-20", "21-30", "31-40", "41-50", "51-60", "61-70", "71-80", "81-90", "91-100")`  
`barplot(marks_AE1,xlab = "Exam mark groups", ylab = "Number of exams", main = "Exam marks for AE 1",col = "Red",names = axis, ylim = c(0,40)) [1]`  
`barplot(marks_AE2,xlab = "Exam mark groups", ylab = "Number of exams", main = "Exam marks for AE 2",col = "Purple",names = axis, ylim = c(0,50))`
- (2)



(1)  
(1)



(4)

- ii) The distributions of marks look similar, especially for middle scores. However, there appears to be some differences in marking for low and high scoring exams.

The plot for associate examiner 1 resembles a Normal shape (but it is not as clear for associate examiner 2, where there appears to be some skewness).

Overall, the plots suggest that the two associate examiners are generally consistent

(3)

- iii)  $H_0$  : difference in means is zero  $v$   $H_1$  : difference in means is not zero.

(1)

```
AE_1 = c(4,1,5,1,4,6,4,5,3,6)
> AE_2 = c(3,2,4,0,3,4,2,3,3,6)
> t.test(AE_1,AE_2,paired=TRUE,var.equal = TRUE)
```

```
Paired t-test
data: AE_1 and AE_2
t = 2.862, df = 9, p-value = 0.01872
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1886284 1.6113716
sample estimates:
mean of the differences
      0.9
```

(2)

As the p-value is 0.01872 which is lower than 5%, we have sufficient evidence to reject the null hypothesis and hence we can conclude that there is difference between the two markets at 5% level of significance.

(1)

(4)

- iv)  $H_0$  : variance of marks given by two associate examiners is equal  $v$   
 $H_1$  : variance of marks given by two associate examiners is not equal.

(1)

```
> var.test(x=AE_1,y=AE_2,level=0.95)
```

F test to compare two variances

```
data: AE_1 and AE_2
```

(2)

F = 1.3136, num df = 9, denom df = 9, p-value = 0.6911  
 alternative hypothesis: true ratio of variances is not equal to 1  
 95 percent confidence interval:  
 0.3262887 5.2886923  
 sample estimates:  
 ratio of variances  
 1.313636

As the p-value is 0.6911 which is greater than 5%, we have insufficient evidence to reject the null hypothesis and hence we can conclude that at 5% level of significance, the assumption of equal variances holds true.

(1)

(4)

- v) Since, the two papers are different; we cannot do a paired test now. We will do a t-test with paired=FALSE.

```
> AE_1_First_Paper<-c(4,1,5,1,4,6,4,5,3,6)
> AE_2_Second_Paper<- c(3,2,4,0,3,4,2,3,3,6)
>
> t.test(AE_1_First_Paper,AE_2_Second_Paper,paired=FALSE,var.equal = TRUE)
```

#### Two Sample t-test

data: AE\_1\_First\_Paper and AE\_2\_Second\_Paper  
 t = 1.1968, df = 18, p-value = 0.2469  
 alternative hypothesis: true difference in means is not equal to 0  
 95 percent confidence interval:  
 -0.6799654 2.4799654  
 sample estimates:  
 mean of x mean of y  
 3.9 3.0

Required confidence interval is [-0.6799654 , 2.4799654]

(2)

- vi) Since 0 lies in the 95% confidence interval, we conclude that for (v), there is insufficient evidence to reject the null hypothesis at 5% level of significance and hence there is no difference in the marking of the two associate examiners.

The same test when performed as a paired t-test, it gave a result that there is difference in the marking of the two examiners.

However, when performed as a t-test between two independent samples, the conclusion of the test was that there is no difference in the marking of the two examiners.

(3)

[20]

#### Solution to Question 4:

- i) The parameters of the posterior distribution are:

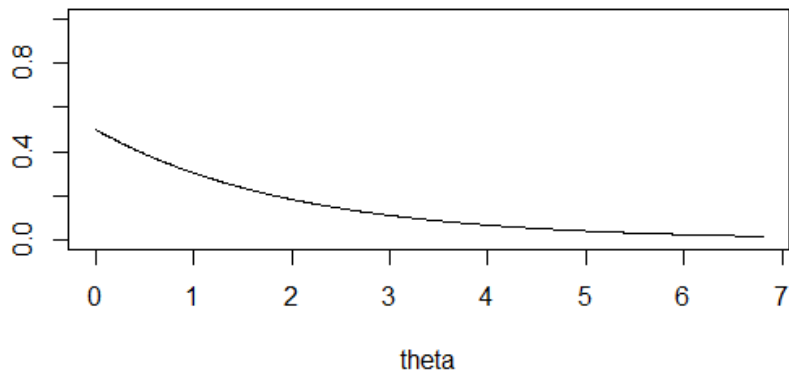
$$\sum_{i=1}^n y + 1 \text{ and } n + \alpha \quad (2)$$

- ii)

theta = seq(0,6.8,by=0.01)

(3)

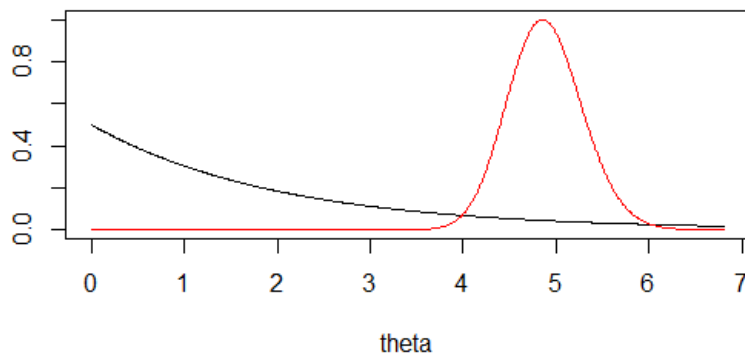
```
plot(theta,dexp(theta,0.5),type="l",ylab="", ylim=c(0,1),xlab="theta")
```



iii)

```
y = c(5,5,6,2,4,10,2,5,5,2,5,3,7,4,4,5,4,6,7,2,8,4,6,4,3, 6,6,6,5,7)
```

```
lines(theta,dgamma(theta,sum(y)+1,30+0.5),col="red")
```



(4)

iv) Whereas the prior density is a downward sloping curve, the posterior distribution has a jump between  $\theta = 3.8$  to  $6$ . This indicates that the posterior distribution is more affected by the underlying data of  $Y$  (30 observations) rather than the prior distribution.

(2)

```
v) qgamma(c(0.05,0.95),sum(y)+1,30+0.5)
```

```
[1] 4.246111 5.561647
```

So, 90% equal-tailed credible interval using the posterior distribution is

```
(4.246111, 5.561647)
```

(2)

```
vi) 1-pgamma(5,sum(y)+1,30+0.5)
```

```
[1] 0.3776056
```

(2)

[15]

\*\*\*\*\*